SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

Application of Data Mining Techniques in Improving Breast Cancer Diagnosis

#SASGF

APPLICATION OF DATA MINING TECHNIQUES IN IMPROVING BREAST CANCER DIAGNOSIS

Josephine S Akosa* & Shannon Kelly**

*PhD in Statistics, Oklahoma State University

**MBA, Marketing in Analytics Certificate, Oklahoma State University

ABSTRACT

Breast cancer is the second leading cause of cancer deaths among women in the United States. Although mortality rates have been decreasing over the past decade, it is important to continue to make advances in diagnostic procedures as early detection vastly improves chances for survival.

The goal of this study is to identify a data mining model that accurately predicts the presence of a malignant tumor using data from fine needle aspiration (FNA) with visual interpretation. Furthermore, this study aims to identify the variables most closely associated with accurate outcome prediction.

Ultimately, a gradient boosting model utilizing a principal component variable reduction method was selected as the best prediction model with a 2.4% misclassification rate, 96.27% specificity, 100% sensitivity, 0.963 Kolmogorov-Smirnov statistic, 0.985 Gini coefficient, and 0.992 ROC index for the validation data. Additionally, the uniformity of cell shape and size, bare nuclei, and bland chromatin were consistently identified as the most important FNA characteristics across variable selection methods.

DATA DESCRIPTION AND PREPARATION

- The study utilizes the Wisconsin Breast Cancer data, originally compiled by Dr. William H. Wolberg and available within the UCI Machine Learning Repository.
- The dataset contains 699 clinical case samples (65.52% benign and 34.48% malignant) assessing the nuclear features of fine needle aspirates taken from patients' breasts.
- There are 11 attributes per observation including the ID and the binary target variable. The target variable diagnoses whether the tumor is benign (non-cancerous) or malignant (cancerous). The remaining input variables are measured on an ordinal scale (1-10), with 1 indicating a normal state and a value of 10 indicating a highly abnormal state.
- To address the high dimensionality of the categorical variables, the weights of evidence approach (WOE) was used to convert the categorical variables into numerical values after which various variable reduction techniques were employed. The WOE approach was implemented via the INTERACTIVE GROUPING node of SAS Enterprise Miner.
- For WOE approach, consider a binary target Y with levels; 0 and 1, where Y = 1 is the event of interest. Now, consider an input variable X with "m" categories. Then WOE is calculated as

$$WOE_i = \log \frac{P(X = x_i|Y = 1)}{P(X = x_i|Y = 0)} \quad for \;\; i = 1,2,\cdots,m$$

- To ensure honest assessment of the models built, the data was partitioned into training (70%) and validation (30%) subsets.
- Prior probabilities were set to account for oversampling since the data was imbalanced.

DATA DESCRIPTION AND PREPARATION CONTINUED

Variable	Label	Mean	Standard Deviation	Minimum	Median	Maximum	Skewness	Kurtosis
WOE_BC	Uniformity of Cell Size	0.478	2.890	-5.398	0.625	3.685	-0.632	-0.829
WOE_BN	Uniformity of Cell Shape	0.245	2.730	-4.381	2.214	2.214	-0.824	-1.099
WOE_CT	Bland Chromatin	0.071	2.775	-5.737	1.371	3.195	-0.946	-0.131
WOE_MAdh	Bare Nuclei	0.036	2.543	-5.184	1.874	1.874	-1.106	-0.224
WOE_Mit	Clump Thickness	-0.051	1.366	-3.392	0.564	0.564	-1.866	1.665
WOE_NN	Single Epithelial Cell Size	0.054	2.397	-5.415	1.636	1.636	-1.151	-0.155
WOE_SECS	Marginal Adhesion	0.432	2.425	-4.283	2.276	2.276	-0.696	-1.252
WOE_UCSh	Normal Nucleoli	1.254	3.622	-3.844	4.528	4.528	-0.411	-1.582
WOE_UCSz	Mitoses	0.847	3.674	-4.668	3.905	3.905	-0.583	-1.414

Table 1. Weight of evidence variable summary statistics

- Analysis of the summary statistics of the WOE variables (Table 1) does not give any indication that variable transformation is necessary.
- Variable importance is judged by the Gini statistic and information value of the WOE variables (Table 2).

Variable	Label	Gini Statistic	Information Value	Information value ordering
UCSz	Uniformity of Cell Size	95.155	6.786	1
UCSh	Uniformity of Cell Shape	94.597	6.529	2
BC	Bland Chromatin	87.958	4.770	3
BN	Bare Nuclei	86.662	4.755	4
CT	Clump Thickness	80.325	4.190	5
SECS	Single Epithelial Cell Size	85.119	4.090	6
MAdh	Marginal Adhesion	80.286	3.954	7
NN	Normal Nucleoli	78.203	3.823	8
Mit	Mitoses	42.318	1.507	9

Table 2. Variable importance of the Weights of evidence variables

APPLICATION OF DATA MINING TECHNIQUES IN IMPROVING BREAST CANCER DIAGNOSIS

Josephine S Akosa* & Shannon Kelly**

*PhD in Statistics, Oklahoma State University

**MBA, Marketing in Analytics Certificate, Oklahoma State University

METHODS

- Prior to model building, several variable selection/reduction nodes in SAS Enterprise Miner were implemented to select the most significant input variables, including: variable selection, variable clustering, decision tree, partial least squares, principal component analysis, regression and LARS.
- A variety of data mining techniques were considered for model building (Fig. 1).
- All models were built in SAS Enterprise Miner 13.1.

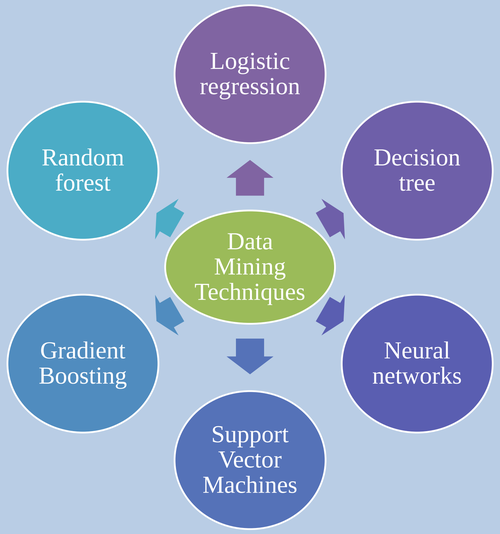

Fig. 1. Data Mining Techniques employed

Models built included:

- Logistic regression with variation in variable selection criteria (default, stepwise, backward, decision tree, principal components)
- Decision tree with variation in splitting rule target criteria (default, entropy, Gini, number of branches
- Neural and autoneural network via variable selection
- Gradient boosting via variable selection
- Random forest via variable selection
- Support Vector machine with variation in kernel function (linear, sigmoid, polynomial) via variable selection

Model Selection criterion included the following validation metrics:

- Misclassification rate
- Specificity
- Sensitivity
- Kolmogorov-Smirnov statistic
- Gini coefficient
- ROC index

RESULTS

- In selecting the best model, the misclassification rate was given the highest importance followed closely by the sensitivity and specificity rates.
- After comparing all the models, the gradient boosting model via principal components (Boosting via PC) was selected as the best model (Table 3). When compared to the other models, the selected model has the highest sensitivity and KS statistic, lowest misclassification rate and the second highest specificity, Gini coefficient and ROC index.
- The decision tree and the random forest via principal components also provide a relatively good model for outcome prediction as can be seen by the validation sensitivity, specificity and misclassification rates.
- Among the variable reduction techniques, the principal components were the most significant variables in reducing the model comparison fit statistics.

Model Description	Misclassification rate	KS Statistic	Gini Coefficient	ROC Index	Sensitivity	Specificity
Boosting via PC	0.024	0.963	0.985	0.992	100.00%	96.27%
Decision tree via PC	0.029	0.949	0.949	0.974	98.63%	96.27%
Random Forest via PC	0.029	0.949	0.969	0.984	98.63%	96.27%
Autoneural via regression	0.029	0.943	0.979	0.990	97.26%	97.01%
Linear Logistic regression	0.034	0.940	0.982	0.991	97.26%	96.27%
Random Forest via regression	0.034	0.940	0.982	0.991	97.26%	96.27%
Random Forest via PLS	0.034	0.935	0.977	0.989	97.26%	96.27%
Autoneural (default)	0.034	0.963	0.988	0.994	95.89%	97.01%
Decision tree (3 branches)	0.043	0.920	0.937	0.968	97.26%	94.78%
SVM (Linear)	0.043	0.942	0.984	0.992	95.89%	95.52%

Table 3. Model comparison fit statistics

Explaining the best model:

- Gradient boosting models combine predictions from a set of decision trees into a single prediction model with the ultimate goal of increasing the probability of selecting an observation that aids in predicting the target variable accurately.
- The technique builds a series of incrementally improved decision trees through resampling of the data set with replacement to produce results that form the weighted average of the resampled data.
- The algorithm places greater weights on misclassified cases as the model develops.

APPLICATION OF DATA MINING TECHNIQUES IN IMPROVING BREAST CANCER DIAGNOSIS

Josephine S Akosa* & Shannon Kelly**

*PhD in Statistics, Oklahoma State University

**MBA, Marketing in Analytics Certificate, Oklahoma State University

RESULTS CONTINUED

- With regards to the selected gradient boosting model, the first 5 principal components (PC) were used in the model building. These components account for 90.48% of the total variability in the data.
- Additionally, the first PC was identified as the most important variable with 20 splitting rules and a value of 1 for the variable importance.

Variable	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8
WOE_UCSz	0.939	-0.091	0.000	-0.052	-0.007	-0.098	-0.115	-0.170
WOE_UCSh	0.911	-0.150	0.042	-0.045	-0.087	-0.147	-0.105	-0.267
WOE_SECS	0.885	-0.036	-0.068	-0.106	0.022	-0.192	-0.244	0.320
WOE_BC	0.857	-0.144	-0.081	-0.050	0.021	0.463	-0.145	0.009
WOE_NN	0.834	0.074	-0.143	-0.407	-0.128	0.004	0.309	0.041
WOE_BN	0.829	-0.153	0.018	0.399	-0.317	0.003	0.139	0.097
WOE_MAdh	0.828	-0.061	-0.307	0.212	0.376	-0.055	0.163	-0.016
WOE_CT	0.750	0.017	0.626	-0.010	0.173	0.024	0.105	0.051
WOE_Mit	0.632	0.761	-0.031	0.108	-0.038	0.033	-0.065	-0.043

Table 4. Correlation between observed variables and principal components

- Principal component analysis (PCA) is a technique used to convert a set of potentially correlated observations into sets of uncorrelated variables.
- The first PC accounts for majority of the total variance within the variables. As a result, this component will be correlated with at least some of the observed variables.
- The correlation of the WOE of the observed variables and the principal components are displayed in Table 4. In this analysis, a correlation value of 0.5 in absolute value is deemed significant.

CONCLUSIONS

- The gradient boosting model turned out to be the best model for diagnosing breast cancer using data from fine needle aspiration.
- Uniformity of cell shape and size, bare nuclei, and bland chromatin were identified as the best FNA characteristics with respect to breast cancer diagnosis.
- These results indicate that outcome prediction can be further improved by refining the methods used to identify and measure the FNA characteristics. For example, technological advances that improve the reliability of uniformity estimates could improve the results of the data mining models.
- Finally, utilizing this model would help decrease interpretation errors by radiologists.
- In order to validate these findings, it is important for further research to be conducted; including applying this method to other types of malignant tumor diagnosis.

ACKNOWLEDGEMENT

We wish to express our sincere gratitude to Dr. Goutam Chakraborty, Department of Marketing and founder of SAS and OSU Data Mining Certificate program – Oklahoma State University for his support and guidance throughout this study.

CONTACT INFORMATION

Josephine Sarpong Akosa
Department of Statistics
Oklahoma State University
320-C Math Sciences (MSCS)
Stillwater, OK 74078-1056
josephine.akosa@okstate.edu
915-407-3650

REFERENCES

- American Cancer Society. Cancer Facts & Figures 2015. Atlanta: American Cancer Society; 2015
- Breast Cancer Wisconsin (Original) Data Set (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original))
- Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, *34*(2), 113-127
- Fine Needle Aspiration (Fine Needle Biopsy) (http://ww5.komen.org/BreastCancer/FineNeedleBiopsy.html)
- Principal Component Analysis (http://support.sas.com/publishing/pubcat/chaps/55129.pdf)
- Saarenmaa, I., Salminen, T., Geiger, U., Heikkinen, P., Hyvärinen, S., Isola, J., ... & Hakama, M. (2001). The effect of age and density of the breast on the sensitivity of breast cancer diagnostic by mammography and ultrasonography. *Breast cancer research and treatment*, *67*(2), 117-123.

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

LAS VEGAS | APRIL 18-21

#SASGF